Proposal: A New Method for Estimating Precision and Bias for Gaseous Automated Methods for the Ambient Air Monitoring Program

Executive Summary

In order to provide decision makers with data of adequate quality, OAQPS is using the DQO process to determine our data quality needs for our ambient air criteria pollutants. There are some data quality indicators that drive the DQOs that relate directly to the measurement system being used to collect the ambient air measurements. These data quality indicators include precision, bias, completeness and sampling frequency. These variables need to be in certain acceptable ranges (called measurement quality objectives) in order for us to make decisions with specified levels of confidence.

As was written in the document entitled *"Guideline on the Meaning and Use of Precision and Accuracy Data Required by 40 CFR Part 58 Appendices A and B*", " the P&A statistics represented a compromise between (a) theoretical statistical exactness, and (b) simplicity and uniformity in computational procedures". The current CFR Appendix A statistics (with the exception of $PM_{2.5}$) aggregate precision and bias into one statistic which make it difficult to use for input to the DQO process.

OAQPS is recommending developing the precision and bias measurement quality objectives on confidence intervals at the site level of data aggregation. Since the criteria pollutant data are used for very important decisions (comparison to the NAAQS) it is felt that providing precision and bias estimates at upper confidence limits provides a higher probability of making appropriate decisions. This statistic provides a conservative approach to measuring precision and bias. The intent of this is to move organizations to a "performance based" quality system; allowing organizations that show tight acceptable results the flexibility in reducing the frequency of certain QC checks and the ability to focus their quality system resources where it will do the most good.

Estimates of both bias and precision for the four automated gaseous methods (O_3, CO, SO_2, NO_2) will be derived from the bi-weekly "precision" checks. Since every site performs the precision checks at an acceptable frequency we have enough information to assess and control data quality at this level.

A focus group made up of OAQPS, EPA Regions and monitoring organizations reviewed three statistical estimates of ozone precision and bias by means of a number of data models and real concentration data from AQS. The focus group recommends using an absolute bias confidence interval and the signed CV confidence interval as the statistic for setting the acceptable measurement quality objectives for the data quality indicators of precision and bias.

1.0 Introduction

In order to provide decision makers with data of adequate quality, OAQPS is using the DQO process to determine our data quality needs for our ambient air criteria pollutants. There are some data quality indicators that drive the DQOs that relate directly to the measurement system being used to collect the ambient air measurements. These data quality indicators include precision, bias, completeness and sampling frequency (the more samples generally the better). These variables need to be in certain acceptable ranges (measurement quality objectives) in order for us to make decisions with specified levels of confidence.

We think two things should be described in CFR. One is the statistics used to estimate these variables and second is the measurement quality objectives for each variable. This is different from what is currently in CFR. The current CFR (except for $PM_{2.5}$) is focused more on quality control information and addresses questions like whether each of the precision checks in a quarter for each site is in control and whether all of the precision checks in a quarter for each reporting organization are tight. These are good questions to ask and they help to quality control the ambient air monitoring networks. However, these are not the questions for which the networks are designed. They are designed to answer regulatory questions and we think CFR needs to include the information about how we know if the networks are providing data of sufficient quality to answer these regulatory questions.

As was written in the document entitled *"Guideline on the Meaning and Use of Precision and Accuracy Data Required by 40 CFR Part 58 Appendices A and B*", " the P&A statistics represented a compromise between (a) theoretical statistical exactness, and (b) simplicity and uniformity in computational procedures". The current CFR Appendix A statistics (with the exception of $PM_{2.5)}$ lump precision and bias together. In developing the statistics used to estimate the variables used in the DQO process, we are examining ways that are statistically reasonable and simple and that are based on the measurements that we currently collect in the ambient air QA/QC programs. We are also looking for consistency in the statistics, if possible. The P&A statistics for PM2.5 currently differ from the P&A statistics for all of the other methods and the P&A statistics for the manual methods differ from those for the automated methods. Hopefully, so many different statistics are not necessary and we will be able to use one set or maybe two sets of statistics for all the criteria pollutants.

Over the last two years OAQPS has been working with the QA Strategy Workgroup to review and improve the Ambient Air Monitoring Program's quality system. As part of this process the Workgroup has endorsed the use of the DQO process and has also been reviewing the measurement quality objectives that lead to attainment of the DQOs. A focus group from the larger QA Strategy Workgroup was formed to work through some alternative methods to estimate precision and bias for the automated methods (CO, NO_2, SO_2 and O_3). The following information represents the activity and the proposed statistics for these estimates.

2.0 Discussion

OAQPS is recommending developing acceptance criteria based on confidence intervals. That is, determining whether the bias and precision variables meet the measurement quality objectives will be based on whether the confidence intervals for these variables meet the measurement quality objectives. The intent of this is two-fold. One reason for using confidence intervals is to be confident the measurement quality objectives are being met. We believe it is different to say the bias is 5% plus or minus 10% compared to saying the bias is 5% plus or minus 1%. A second, and very practical, reason for using confidence intervals is to allow organizations that show tight acceptable results the flexibility in reducing the frequency of certain QC checks. For example, the site with a bias of 5% plus or minus 1% likely does not need as many QC checks as the site with the bias of 5% plus or minus 10%. The acceptance criteria would also be based on the number of years of data that coincide with the time frame of the ambient air quality standards. For example, since the 8-hour ozone standard is based on 3 years of data, the acceptance criteria for bias and precision will also be based on 3 years of data. Additionally, the acceptance criteria apply to each site operating an automated method.

As mentioned above, we are hoping to use existing measurements to estimate bias and precision. For the automated methods, we believe we can estimate both bias and precision from the bi-weekly "precision" checks (henceforth called bi-weekly checks) and then double-check these bias and precision estimates with the annual accuracy audits, independent State audits and the NPAP Program. To test the reasonableness of estimating bias and precision from bi-weekly checks, we made up some actual/indicated pairs, assuming different sizes of bias and precision, and tested a couple of proposed statistics. We simulated 3 years of data and provided summary statistics at the quarterly, annual, and 3-year level. The reason for the 3-year level is because this is the level to which some of our standards ($PM_{2.5}$ and ozone) aggregate. Remember, we are thinking of summary statistics to support regulatory decisions and regulatory decisions are usually based on multiple years of data. For each scenario, we summarized the data by the three methods below. The statistical equations are provided in Attachment A.

1. **CFR Probability Interval**. For these statistics, we show what is currently in CFR, namely the overall percent difference in the actual and indicated concentrations and an associated probability interval that shows where 95% of all the percent differences should fall. Note that this does not provide separate estimates of bias and precision.

2. **Signed Bias & Precision (CV)**. For this case, we estimate bias and precision separately. We also estimate confidence intervals for bias and confidence intervals for precision. A comment on this approach is that if there is trend in bias, such as +10% one year, 0% the next year, and –10% the third year, then, from a 3-year perspective, you may say the system is unbiased but very variable. This is how these statistics treat the trend in bias. Thus the bias tends to be small and the precision large, in general.

3. **Absolute Bias & Precision (CV)**. As with the signed case above, we estimate bias and precision separately. We also estimate confidence intervals for bias and confidence intervals for precision. However, if there is trend in bias, such as +10% one year, 0% the next year, and –10% the third year, then, from a 3-year perspective, you may say the system has a great potential for bias but is precise. This is how these statistics treat the trend in bias. Thus the bias tends to be large and the precision small, in general.

Figure 1 shows the results for one of the hypothetical cases studied. It provides an example of the various precision and bias estimates for a 3 year data set where the **true** measurement imprecision is 5% and the true bias is 15% for year 1, 0% for year 2 and -10% for year 3. There are 5 sections to the Figure.

- The left-most area of Figure 1 shows the spread of the relative differences of the biweekly checks for each of the years and for all the years combined. These box and whisker plots show that the bias varies from year to year and that it is decreasing (the center of boxes shifts from around 15% to 0% to –10%) but that the imprecision is small (the boxes are small and whiskers short). On the other hand, the 3-year box and whisker plot shows no bias (the box is centered about 0) but large variation (the box is wide and the whiskers are long).

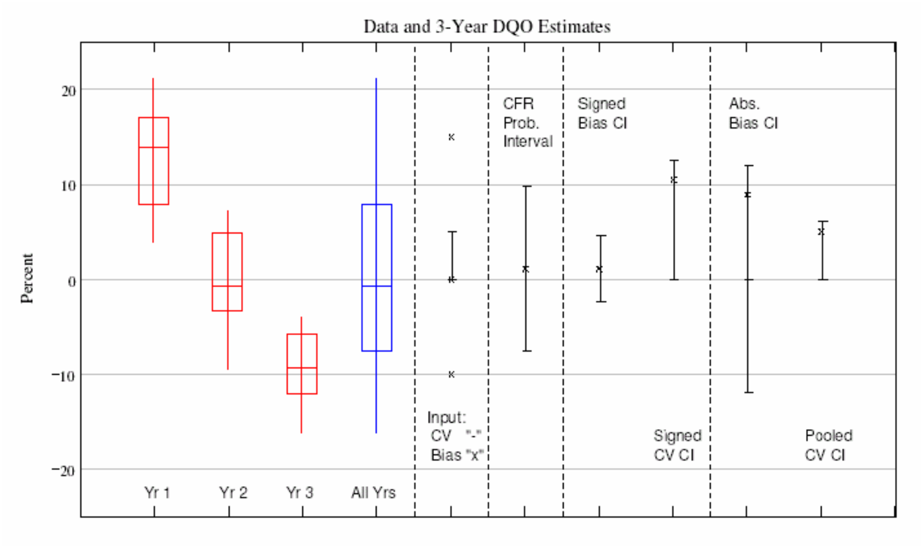
- The next section of Figure 1 shows the true bias (represented as *'s) and imprecision (represented as a line from 0 to the amount of imprecision, 5% in this case).

- The next section shows the results based on the statistics currently in CFR. The center of the interval is represented by the "*." The interval indicates where 95% of the past, present, and future relative differences from the biweekly checks are expected to be.

- The next section shows the bias and precision estimates and their respective confidence intervals for the "Signed" case described above. Precision is represented by the term "CV". The estimates are represented by a "*." Thus bias is nearly 0 and imprecision is around 10%. The confidence interval for the signed bias is always centered on the "*." In this example, the 90% confidence interval for bias is about –3% to 5%. The confidence interval for the precision estimate always extends from 0 to the upper confidence limit. That is, we are not showing the lower confidence limit for precision since it will always be between 0 and the estimate for precision. The upper confidence limit for precision is about 12%. So based on "Signed" estimates, we would say that this site is operating with a bias that is somewhere between –3 and +5% with an imprecision that may be as large as 12%.

- The last section shows the bias and precision estimates and their respective confidence intervals for the "Absolute" case described above. Again, the estimates are represented by a "*" so for this case the bias is about 8% and the imprecision (CV) is around 5%. The confidence interval for the absolute bias is always centered on 0. So for this example, we are saying that the bias has the potential to be as large as +12% or as small as -12% and +12%. And as above, the confidence interval for the precision estimate always extends from 0 the upper confidence limit, which is about 6% in this case. So for the "Absolute" estimates, we would say that this site is operating with a potential bias as small as –12% or as large as 12% with an imprecision that may be as large as 6%.

Figure 1. Precision and Bias Estimates for a Hypothetical Example



We reviewed seven case studies with varying true precisions and biases to understand how the different types of statistics "reacted". Next we looked at some real examples based on 3 years of biweekly checks of ozone during 1999 to 2001. We looked at a "typical" and an "extreme" reporting organization. For the "typical" reporting organization, the percent differences from the biweekly checks are approximately the same for each site and year and the percent differences are tight. See Figure 2. For the "extreme" reporting organization, the percent differences from the biweekly checks are more variable both within a site and from site to site. See Figure 3. The second case is rare, but gives an indication of how extreme the ozone relative differences can be.

These two examples were generated to complement the seven hypothetical case studies previously reviewed. The seven case studies were based on "made up" data and, upon inspection, present challenges to the summary statistics. Our goal with these two real case studies is to get a better idea of what we most likely will see in the "real world."

For each of the real case studies, this is what is presented in Figures 2 and 3:

(1) A plot of the spread in the relative differences in the actual and indicated concentrations of ozone based on the biweekly checks. For each site and each quarter, there is a box and whisker plot, where the whiskers extend from the minimum to the maximum relative differences. This plot is hard to read, but it gives a general sense of where the relative differences lie. A different color is used for each site.

(2) More box and whisker plots of the relative differences. These show how much the relative differences vary for each site. That is, we throw together all the biweekly checks for each site and show where the bulk of the data are and how spread it is.

(3) To the right of the second set of box and whiskers is a small table. The table provides

some of the summary statistics for each site. It tells the average percent difference and spread in the percent differences for each site. These are just like the statistics as currently calculated in CFR for the site-level aggregation.

(4) In the lower part of each page is information about various "intervals." The first interval is the probability interval, as given in CFR, for the reporting organization. The next intervals break the relative differences into bias and precision components. The first set is based on a conservative approach that tends to estimate a large bias and smaller precision. The second set tends to estimate smaller bias and larger precision. The details of the statistics are given in Attachment A.

Some attributes of these 2 cases are

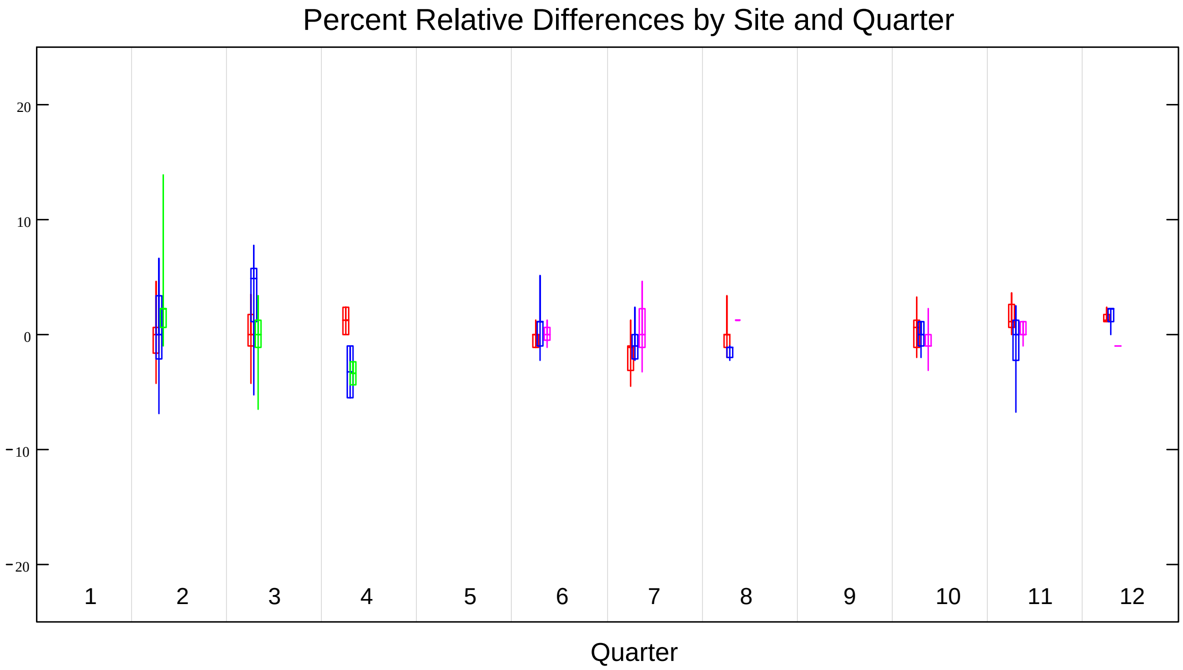
"Typical RO"

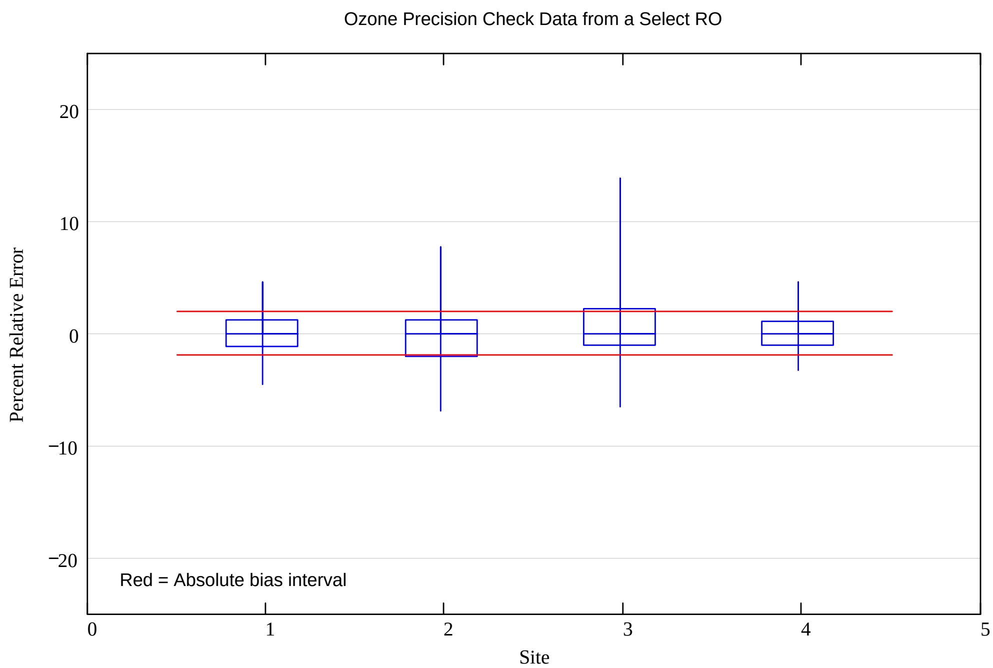
- relative percents generally tight
- each site doing about what every other site doing
- CFR says 95% of all checks have been/will be in -5% to 5%
- approach that emphasizes bias (Abs_Bias), says bias likely in the -2% to 2% range and the precision about 2.7%
- approach that emphasizes precision (Sign_Bias), says bias is about zero and that the precision is about 2.7%

"Variable RO"

- relative percents have more variability, AND the sites seem to differ from each other, look at quarters 8, 11, 12
- looking at each site, we see that site 7 looks quite different from the rest of the pack; it tends to have negative percent differences whereas the other sites tend to have positive percent differences; the site-level box and whiskers also show that some sites have more repeatable percent differences as indicated by small boxes and short whiskers (e.g. sites 5 and 6) compared to others (e.g. sites 1, 4, and 11)
- CFR says 95% of all checks in -10% to 13%.
- approach emphasizing bias (Abs_Bias) says bias likely in the -5% to 5% range, tending more to the positive 5% side, though; precision is around 5.2%
- approach emphasizing precision (Sign_Bias) says bias likely in the 1.5% to 2.4% range and precision is around 6.1%.

Figure 2. Percent Differences based on Biweekly Checks of Ozone from a Typical Reporting Organization.





Site_Bias_CV =

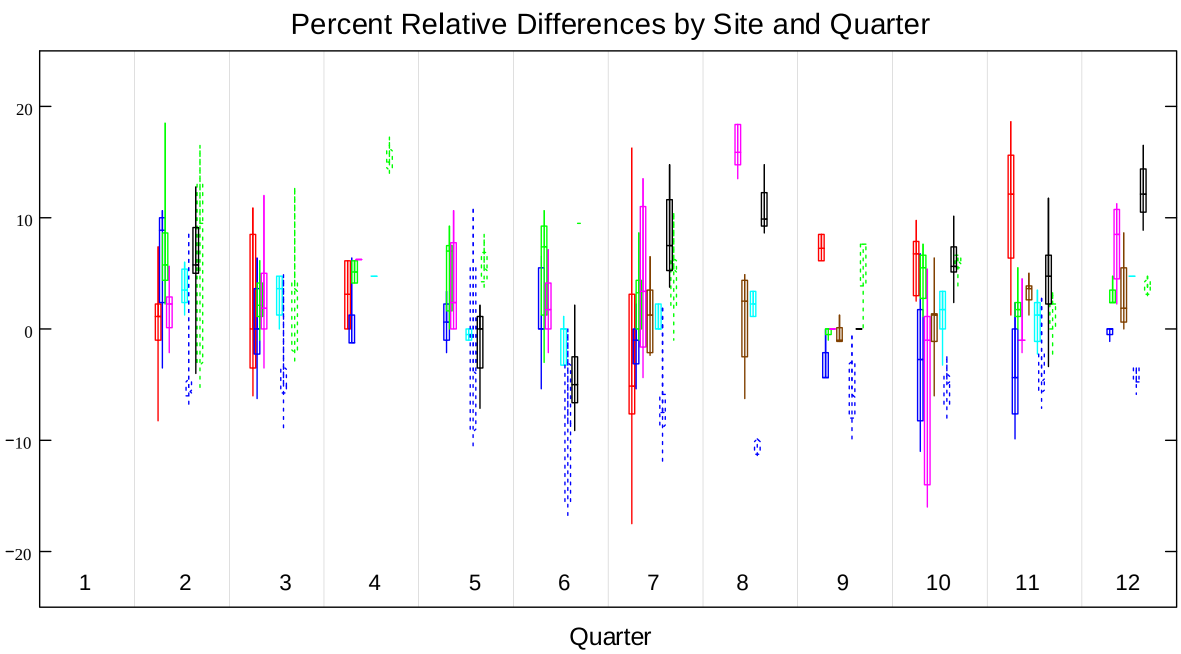
1	-0.1	2
2	0.1	3
3	0.6	4.2
4	-0.1	1.5

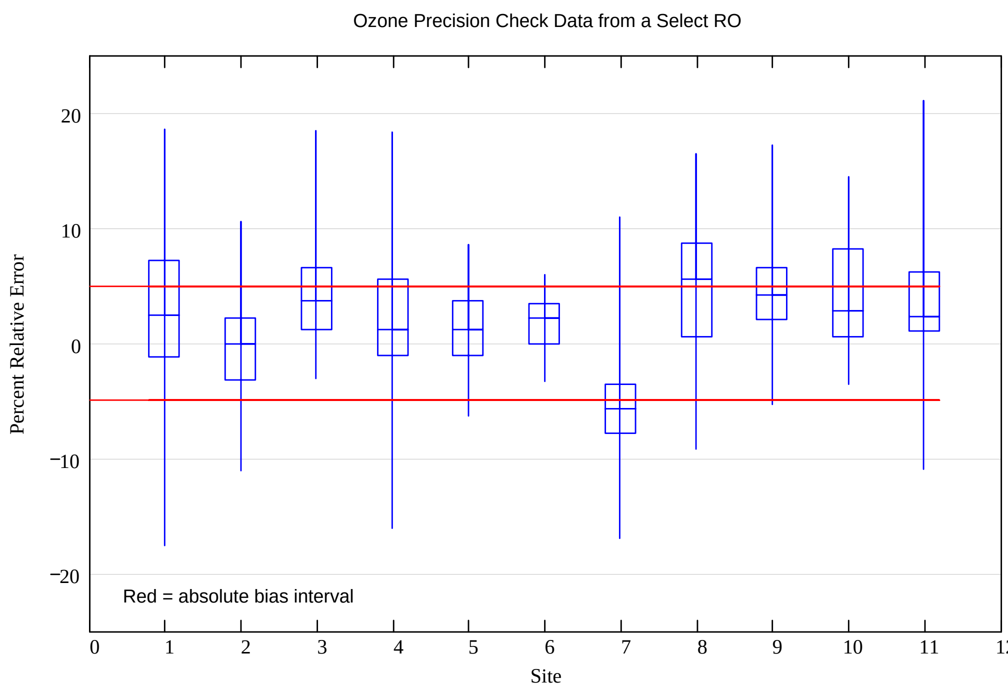
CFR_Prob_Interval = (−4.89 4.96)

Abs_bias_interval = (−1.94 1.94)   Pooled_by_site_CV_UB = 2.7

Sign_bias_interval = (−0.31 0.38)   Signed_CV_UB = 2.7

Figure 3. Percent Differences based on Biweekly Checks of Ozone from a Highly Variable Reporting Organization.





Site_Bias_CV =

1	2.7	7
2	-0.4	4.5
3	4.2	3.7
4	2.6	6.2
5	1.4	3.2
6	1.7	2.5
7	-5.6	4.7
8	4.6	6.3
9	4.8	5.2
10	4	4.4
11	3.7	5.8

CFR_Prob_Interval = (−9.59   13.48)

Abs_bias_interval = (−4.94   4.94)   Pooled_by_site_CV_UB = 5.2

Sign_bias_interval = (1.51   2.38)   Signed_CV_UB = 6.1

3.0 Sample Size Flexibility

Changing from point estimates of the bias and precision data quality indicators to confidence intervals based on upper bounds is more strict, but in spite of being more conservative it is also more flexible in that it the upper bounds are dependent on the total number of biweekly checks. In some cases fewer total checks may be needed than have been traditionally collected.

For ozone and other gasses, the proposed precision and bias estimates are both made from the biweekly checks. Table 1 shows how many of those checks are needed to confidently (90%) establish that both the precision and bias are less than 10%. In this way one knows that both the precision and the bias are controlled to at most 10% provided the sample size is at least as many as shown in Table 1. For Table 1, one-sided 90% confidence limits were assumed. This matches the current use for the $PM_{2.5}$ precision estimates in CFR.

Table 1. Conservative Number of Precision and Bias Checks Needed to Yield Both an Absolute Bias Upper Bound of at Most 10% and an Upper bound of at most 10% for the Precision.

Minimum sample size		Precision Point Estimate				
		5%	6%	7%	8%	9%
Bias Pt. Est.	5%	8	8	12	24	87
	6%	12	12	12	24	87
	7%	20	20	20	24	87
	8%	43	43	43	43	87
	9%	166	166	166	166	166

As an example, let's use site 7 in Case 2 (highly variable reporting organization). This site has a 3 –year bias point estimate of -5.6 (round to 6) and a measurement CV point estimate of 4.7 (round to 5). Using the intersection of the bias and precision information in Table 1 says that one would need 12 biweekly checks to ensure that 90% of the time the upper confidence interval for both bias and precision would be 10% or less. Note that approximately 36 biweekly checks are currently collected for each ozone site (assuming the ozone season for the site is lasts about 2 quarters each year).

The sample size calculations are based on normal distribution theory. In particular, it is being assumed that the relative errors, d, and the absolute relative errors, |d|, are normally distributed. Strictly speaking this cannot be true. However, there are several aspects that justify the use of this theory as an approximation. First, provided that the d's are symmetrically distributed, then the distribution of mean will quickly converge to the normal distribution, making part of the calculations a good approximation. While expect the d's to be symmetrically distributed, taking absolute values will disturb that symmetry and hence diminish the accuracy of the normal distribution approximation for the mean. This loss in accuracy is compensated by using a conservative estimate for the variance. The variance of the relative differences is used rather than the variance of the absolute value of the relative differences ($Var(|d|) \leq Var(d)$). Hence, the calculation of the needed sample size is slightly increased. Finally, the Bonferoni method for multiple comparisons was used to distribute the type 1 error. This is a conservative method which will also protect against any loss in accuracy of the normal distribution theory approximation.

4.0 Conclusions

As the focus workgroup reviewed the various cases and discussed the information we came to the following conclusions:

- **Use the absolute bias confidence interval and the signed precision (CV) confidence interval** as the statistics for setting the acceptable measurement quality objectives for the data quality indicators of precision and bias. This provides a conservative approach but it also allows flexibility in implementing quality control activities and would help focus QA resources where they were most needed. Using case 2 as an example; one could see that a reporting organization might want to focus on reducing the variability at sites 1, 2, 7 and 11 and trying to determine what might be causing the biases at sites 7 and 8. Therefore, one could increase the number of biweekly checks for these sites while reducing the frequency on others that appear to be in tighter control (such as sites 5 and 6).

- **Develop measurement quality objectives at the site level of data aggregation**. Since every site performs the biweekly checks at an acceptable frequency (as determined by the sample size in Table 1) we have enough information to assess and control data quality at the site level. Data will still be presented by reporting organization because several QA decisions and QA implementation occur at this level.

Attachment A

Precision and Bias Statistics

This attachment describes the various precision and bias statistics that were considered during the review of the statistics in 40 CFR Part 58 Appendix A. The basic measure used in all of the following calculations is

$$d_i = \frac{ind_i - act_i}{act_i} \cdot 100 \, ,$$

where ind_i is the concentration as indicated on the analyzer and act_i is the concentration of the audit gas for day i.

1. CFR probability interval

The statistics currently in CFR provide a measure of centrality, which is the overall mean of the d_i's, a measure of spread, which is the pooled standard deviation, and a 95% probability interval, which is an interval indicating where 95% of the past, present, and future d_i's are expected to lie. The probability interval is computed at the reporting organization level for each quarter, meaning that the probability interval is where 95% of the d_i's from all the sites in the reporting organization for the quarter are expected to lie. Statements are not made at the site level.

The equation for the probability interval is

$$\left(m - 1.96 \cdot S, m + 1.96 \cdot S\right)$$

where $m = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i$, $S = \sqrt{\frac{\sum_{k=1}^{a} (n_k - 1) \cdot (S_k)^2}{\sum_{k=1}^{a} (n_k - 1)}}$, $S_k = \sqrt{\frac{n_k \cdot \sum_{i=1}^{n_k} d_i^{\,2} - \left(\sum_{i=1}^{n_k} d_i\right)^2}{(n_k - 1) n_k}}$ for k=1,2,3,…, a, a is the number of units (analyzers) in the reporting organization being pooled, and n is the number of relative percent differences for the quarter in the reporting organization.

2. Bias and precision estimates where the sign of the d_i's is retained

The statistics for the case where the sign of the d_i's is retained provide a measure of centrality (called bias), which as with CFR is the overall mean of the d_i's, a measure of spread (called precision), which is the standard deviation of the d_i's, a confidence interval for bias and an upper confidence limit for precision. These statistics are calculated at the site level, not the reporting organization level.

The statistics are:

$$\text{Bias, } m = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i$$

$$\text{Precision, } s = \sqrt{\frac{n \cdot \sum_{i=1}^{n} d_i^{\,2} - \left(\sum_{i=1}^{n} d_i \right)^2}{(n-1)n}}$$

where n is the number of d_i's within the aggregation period (for example, quarter or year) for a site. Then the confidence interval for bias is:

$$\left(m - t_{\alpha/2,(n-1)} \cdot s/\sqrt{n} \, , m + t_{\alpha/2,(n-1)} \cdot s/\sqrt{n} \right) \text{ and}$$

the upper bound of the confidence interval for the precision estimate is

$$s \cdot \sqrt{\frac{n-1}{\chi^2_{\alpha,(n-1)}}} \, ,$$

where $t_{\alpha/2,(n-1)}$ is the $\alpha/2$ quantile of Student's t distribution with degrees of freedom $(n-1)$ and $\chi^2_{\alpha,(n-1)}$ is the α quantile of the chi-square distribution with $(n-1)$ degrees of freedom.

3. Bias and precision estimates where the sign of the d_i's is not retained

The statistics for the case where the sign of the d_i's is not retained provide a measure of centrality (called bias), which is the overall mean of the absolute values of the d_i's, a measure of spread (called precision), which is a pooled estimate based on quarterly standard deviations of the d_i's, and upper confidence limits for bias and precision. These statistics are calculated at the site level, not the reporting organization level.

The bias point estimates is:

$$\text{Bias, } m = \frac{1}{n} \cdot \sum_{i=1}^{n} |d_i|$$

where n is the number of d_i's within the aggregation period.

The precision estimate is generated by first computing

$$S_k = \sqrt{\frac{n_k \cdot \sum_{i=1}^{n_k} d_i^{\,2} - \left(\sum_{i=1}^{n_k} d_i \right)^2}{(n_k - 1) n_k}}$$

where n_k is the number of d_i's in quarter k. The pooled estimate of precision is

$$\text{Precision (pooled) estimate} = \sqrt{\frac{\sum_{k=1}^{q} (n_k - 1) \cdot (S_k)^2}{\sum_{k=1}^{q} (n_k - 1)}}$$

where q is the number of quarters and S_k is the precision estimate for quarter k. Then the upper bound of the confidence interval for the absolute bias is

$$m + t_{\alpha,(n-1)} \cdot s / \sqrt{n}$$

where $s = \sqrt{\frac{n \cdot \sum_{i=1}^{n} |d_i|^2 - \left(\sum_{i=1}^{n} |d_i| \right)^2}{(n-1) n}}$ and $t_{\alpha,(n-1)}$ is the α quantile of Student's t distribution with degrees of freedom *(n-1)*. The upper bound for the pooled precision estimate is

$$\text{Precision (pooled) estimate} \cdot \sqrt{\frac{n - q}{\chi^2_{\alpha,(n-q)}}},$$

where $\chi^2_{\alpha,(n-q)}$ is the α quantile of the chi-square distribution with *(n-q)* degrees of freedom.